AD-A105 860 PRINCETON UNIV NJ DEPT OF STATISTICS F/G 12/1
THE SECOND REPRESENTING FUNCTION FOR COMPOUND SITUATIONS.(U)
MAR 81 A BRUCE, D PREGIBON, J W TUKEY DAAG29-79-C-0205
UNCLASSIFIED TR-186-SER-2 ARO-16669.7-M NL

1 OF 1
AD A
105860

END
DATE
FILMED
11-81
DTIC

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 16669.7-M	2. GOVT ACCESSION NO AD-A105 860	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) The Second Representing Function for Compound Situations.		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Andrew Bruce, Daryl Pregibon, John W. Tukey		8. CONTRACT OR GRANT NUMBER(s) DAAG29-79-C-0205
9. PERFORMING ORGANIZATION NAME AND ADDRESS Princeton University Princeton, NJ 08544		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE Mar 81
		13. NUMBER OF PAGES 13
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) ARO		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

LEVEL II

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

TR-186-SER

DTIC SELECTE OCT 20 1981 D H

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper provides a convenient method of computing the low-order moments of compound situation order statistics based on higher order representing function. The explicit derivation of the second representing function is given. The one-wild-Gaussian situation is used to illustrate the method. Tables of one-wild-Gaussian order-statistic moments are displayed for selected sample sizes.

DD FORM 1473, 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

406873

10 81

The Second Representing Function for Compound Situations*

Andrew Bruce[1]
Daryl Pregibon[2]
and
John W. Tukey

Technical Report No. 186, Series 2
Princeton University
Department of Statistics
Princeton, New Jersey 08544

ABSTRACT

A random variable X with distribution function F(x) can be written as $x = R(u)$, where $u = F(x)$ and $R = F^{-1}$. The function $R(u)$ is the (first) representing function of X. For certain selected distributions, this representing function can be easily expressed (e.g. logistic, Cauchy), though in general, approximation or tabulation is required (e.g. Gaussian, slash).

A situation $\{X_i : i=1, \ldots, n\}$ is a collection of independently distributed random variables. If the X_i are identically distributed, the situation is termed <u>simple</u>, otherwise the situation is termed <u>compound</u>. For example,

$$X_i \sim (1-\epsilon)F(x_i) + \epsilon G(x_i)$$

is a simple situation, whereas for $\epsilon = k/n$, $k=0,1,\ldots,n$

$$(1-\epsilon)n \quad X\text{'s} \sim F(x)$$

$$\epsilon n \quad X\text{'s} \sim G(x)$$

is a compound situation.

For simple situations, the low-order moments of the order statistics can be conveniently computed in terms of the (first) representing function of X. For compound situations, a first

representing function is not sufficient for computing these moments. This paper provides a convenient method of computing the low-order moments of compound situation order statistics based on higher order representing function. The explicit derivation of the second representing function is given. The one-wild-Gaussian situation is used to illustrate the method. Tables of one-wild-Gaussian order-statistic moments are displayed for selected sample sizes.

March 6, 1981

The Second Representing Function for Compound Situations*

Andrew Bruce[1]

Daryl Pregibon[2]

and

John W. Tukey

Technical Report No. 186, Series 2

Princeton University

Department of Statistics

Princeton, New Jersey 08544

1. Introduction.

Order statistics play an important role in statistics. Many useful estimators are based on linear combinations of order statistics (or selected subsets thereof). Informal inferential procedures (such as probability plotting) are also based on order statistics. Of particular importance are the low order moments of these quantities, specifically the means, variances, and covariances. Tables of these

*Prepared in connection with research at Princeton University, supported by the Army Research Office (Durham).

1 Present Address: 77 Longview Drive, Princeton, N.J. 08540

2 Present Address: Department of Biostatistics, University of Washington, Seattle 98195.

March 6, 1981

moments exist for many of the commonly used sampling situations. In almost all cases, these situations are "simple", corresponding to a sample of independent and identically distributed random variables. (For an outstanding exception see David, Kennedy, and Knight, 1977.)

In this paper we provide a method of computing low-order moments of order statistics from "compound" situations of the form

$$n-1\ X\text{'s} \sim F(x)$$

$$\text{one } X \sim G(x)\ .$$

The method uses what we call the <u>second</u> representing function of X, namely

$$\frac{\partial}{\partial\epsilon} R_\epsilon(u)\Big|_{\epsilon=0}$$

where $R_\epsilon(u) = F_\epsilon^{-1}(u)$ is the <u>first</u> representing function of X for the simple situation

$$X_i \sim F_\epsilon(X_i) = (1-\epsilon)F(x_i)+\epsilon G(X_i), \quad i=1, \ldots, n\ .$$

We illustrate the method using the one-wild Gaussian compound situation

$$n-1\ X\text{'s} \sim \Phi(x) = \text{Gau}(0,1)$$

$$\text{one } X \sim \Phi(x/10) = \text{Gau}(0,100).$$

This compound situation has been used extensively in studies of robust/resistant estimates of location. The case where

$G(x) = \Phi(x/3)$ was used earlier, and is one of the cases tabulated by David, Kennedy, and Knight (1977). Tables of the low-order moments of the corresponding order statistics are long overdue.

Section 2 describes moment calculations for simple-situation order statistics in terms of the first representing function. Section 3 describes moment calculations for compound situation order statistics in terms of the second representing function. The one-wild-Gaussian compound situation is used to illustrate the method in Section 4.

2. Simple Situations.

Consider an iid sample $\{x_i : i=1,\ldots,n\}$ of random variables with distribution function $F(x)$. Let $v_i = x_{(i)}$ denote the ith order statistic with $v_1 \leq v_2 \leq \ldots \leq v_n$. In contrast to the x's, the v's are neither independent nor identically distributed. Let $H(v_i, v_j)$ denote the joint distribution function of v_i and v_j. The product moment of v_i and v_j $y_j > y_i$ is given by

$$m_{ij} = E(v_i v_j) = \int_{-\infty}^{\infty} \int_{-\infty}^{v_j} v_i v_j \, dH(v_i, v_j)$$

where $dH(v_i, v_j)$ is proportional to

$$F^{i-1}(v_i) f(v_i) [F(v_j) - F(v_i)]^{j-i-1} f(v_j) [1 - F(v_j)]^{n-j} dv_i dv_j \quad . \quad (1)$$

The change of variables $u = F(v)$ is monotone so that

$$u_i = F(v_i) \leq u_j = F(v_j)$$

and the above expression becomes

$$m_{ij} = \int_0^1 \int_0^{u_j} R(u_i)R(u_j)\,dH(u_i,u_j)$$

where $dH(u_i,u_j)$ is proportional to

$$u_i^{i-1}(u_j-u_i)^{j-i-1}\ (1-u_j)^{n-j}\ du_i du_j\ .$$

Thus, given the representing function $v = R(u)$, the low order moments can be obtained, by numerically integrating over the unit triangle $0 \leq u_i \leq u_j \leq 1$. Where the representing function cannot be given explicitly, a numerical approximation to $R(u)$ is required.

* a special form *

Quadrature formulas to obtain an estimate $\hat{m}_{ij}$ of m_{ij} are sometimes more convenient if the region of integration is the unit square rather than the unit triangle, and if integration involves a product form in place of $dH(u_i,u_j)$. This is easily obtained by a further change of variables. Let

$$u_i = (1-z)w$$

$$1-u_j = (1-z)(1-w)$$

where $0 \leq w \leq 1$, $0 \leq z \leq 1$. Then $u_j-u_i = z$ and since the

Jacobian of this transformation is $1-z$, $dF(u_i,u_j)$ is proportional to

$$w^{i-1}(1-w)^{n-j}z^{j-i-1}(1-z)^{n-j+i}\,dwdz \ .$$

That is, w and z are independently distributed as beta random variables:

$$w \sim \beta(i,n-j+1) \text{ and } z \sim \beta(j-i,n-j-i+1) \ .$$

The alternate expression for the product moment of v_i and v_j is therefore

$$m_{ij} = \int_0^1 \int_0^1 R(w(1-z))R(w(1-z)+z)\,d\beta_w(w)\,d\beta_z(z) \ .$$

If desired, one-dimensional quadrature formulas specialized for integrating a function of a beta-variable could now be used, iterating the integral. The accuracy of such quadrature rules has not been explored in detail.

3. Compound Situations.

Consider a realization $\{x_i : i=1,\ldots,n\}$ of random variables from

$$n-k \ X\text{'s} \sim F(x)$$

$$k \ X\text{'s} \sim G(x)$$

for $k=0,\ldots,n$. Let $v_i = x_{(i)}$ denote the ith order statistic with $v_1 \leq v_2 \leq \ldots \leq v_n$. The product moment of v_i and v_j is

$$m_{ij}^{(k)} = E(v_i v_j) = \int_{-\infty}^{\infty} \int_{\infty}^{v_j} v_i v_j \, dH^{(k)}(v_i, v_j)$$

where $H^{(k)}(v_i,v_j)$ is the joint distribution of v_i and v_j. This joint distribution can be derived from that of $v_1, \ldots, v_n$:

$$H^{(k)}(v_1, v_2, \ldots, v_n) = \sum_{\binom{n}{k}} k! \prod_{i \in G} G(v_i)\,(n-k)! \prod_{i \in F} F(v_i) \ .$$

It is easy to see that the resulting formula for $H^{(k)}(v_i,v_j)$ is appreciably more cumbersome than its simple-situation counterpart. Direct integration over $H^{(k)}(v_i,v_j)$ is not particularly attractive, especially if there is a simpler means to attain the same end.

Consider the simple mixture situation

$$X_i \sim F_\alpha(x_i) = (1-\alpha)F(x_i) + \alpha G(x_i) \qquad i = 1, \ldots, n \ .$$

The joint distribution of v_i and v_j is $H_\alpha(v_i,v_j)$ and can be obtained using equation (1). This leads to the simple-mixture-situation order statistic moments:

$$m_{ij}(\alpha) = \int_{-\infty}^{\infty} \int_{-\infty}^{v_j} v_i v_j \, dH_\alpha(v_i, v_j)$$

$$= \int_0^1 \int_0^{u_j} R_\alpha(u_i) R_\alpha(u_j) \, dH(u_i, u_j) \ .$$

where $R_\alpha(u) = F_\alpha^{-1}(u)$ is a first representing function for the mixture.

Alternatively, for any α, we have

$$H_{\alpha}(v_i, v_j) = \sum_{k=0}^{n} \Pr\{k\text{-wild}\} \cdot H^{(k)}(v_i, v_j) \; .$$

$$= \sum_{k=0}^{n} \binom{n}{k} \alpha^{k} (1-\alpha)^{n-k} \; H^{(k)}(v_i, v_j) \; .$$

The simple-mixture-situation order-statistic moments are

$$m_{ij}(\alpha) = \int_{-\infty}^{\infty} \int_{-\infty}^{v_j} v_i v_j \; dH_{\alpha}(v_i, v_j)$$

$$= \sum_{k=0}^{n} \binom{n}{k} \; \alpha^{k} (1-\alpha)^{n-k} \int_{-\infty}^{\infty} \int_{-\infty}^{v_j} v_i v_j \; dH^{k}(v_i, v_j)$$

$$= \sum_{k=0}^{n} \binom{n}{k} \; \alpha^{k} (1-\alpha)^{n-k} \; m_{ij}^{(k)} \; . \tag{2}$$

This fundamental relationship between mixture and k-wild order statistic moments allows the latter to be calculated simply. In particular, for k=1, equation (2) becomes

$$m_{ij}(\alpha) = (1-\alpha)^{n} m_{ij}^{(0)} + n\alpha (1-\alpha)^{n-1} m_{ij}^{(1)} + O(\alpha^{2}) \; .$$

Differentiation with respect to α and evaluation at $\alpha = 0$ leads to

$$\frac{\partial}{\partial \alpha} m_{ij}(\alpha) \Big|_{\alpha=0} = -n \cdot {}_{ij}^{(0)} + n \cdot m_{ij}^{(1)} \; .$$

This implies that the one-wild product moment can be written as a linear combination of the uncontaminated product moment and a term due to the contamination viz

$$m_{ij}^{(1)} = m_{ij}^{(0)} + \frac{1}{n}\frac{\partial}{\partial\epsilon} m_{ij}(\epsilon)\Big|_{\epsilon=0}$$

Algebraically, the correction factor is obtained by differentiating equation (2):

$$\frac{\partial}{\partial\epsilon} m_{ij}(\epsilon) = \int_0^1 \int_0^{u_j} \left[\frac{\partial}{\partial\epsilon} R_\epsilon(u_i) \cdot R_\epsilon(u_j) + R_\epsilon(u_i) \cdot \frac{\partial}{\partial\epsilon} R_\epsilon(u_j)\right] dH(u_i, u_j)$$

and then setting $\epsilon=0$, to give

$$\frac{\partial}{\partial\epsilon} m_{ij}(\epsilon)\Big|_{\epsilon=0} = \int_0^1 \int_0^{u_j} \left[R_1(u_i) R_0(u_j) + R_0(u_i) R_1(u_j)\right] dH(u_i, u_j) \ .$$

In this latter equation $R_0(u)$ is the <u>first</u> representing function of X at $\epsilon=0$ contamination, and $R_1(u)$ is the <u>second</u> representing function defined by $\frac{\partial}{\partial\epsilon} R_\epsilon(u)\Big|_{\epsilon=0}$.

The corresponding compound-situation order-statistic moments are obtained as

$$m_{ij}^{(1)} = m_{ij}^{(0)} + \frac{1}{n} \int_0^1 \int_0^{u_j} \left[R_1(u_i) R_0(u_j) + R_0(u_i) R_1(u_j)\right] dH(u_i, u_j) \ . \qquad (3)$$

(Note that the sample size enters the second term through both n and $dH(u_i, u_j)$.) This expression can be numerically evaluated with little extra effort beyond that for the simple-situation moments $m_{ij}^{(0)}$. Extensions to k-wild compound situations are easily obtained as functions of the representing functions of order up to k+1, where in general

$$a_n = \left(\frac{\partial}{\partial \epsilon}\right)^n R_\epsilon(u)\Big|_{\epsilon=0} .$$

4. The One-Wild-Gaussian Situation.

We now illustrate the preceding discussion using the compound situation

$$n-1 \text{ X's} \sim \Phi(x)$$

$$\text{one X} \sim \Phi(x/10) .$$

To do so, we need an expression for the <u>first</u> representing function $R_\epsilon(u) = \Phi_\epsilon^{-1}(u)$ where

$$u = \Phi_\epsilon(x) = (1-\epsilon)\Phi(x) + \epsilon\Phi(x/10) .$$

Now since $R(\Phi(x)) = x$, we have

$$R(\Phi_\epsilon(x)) = R(\Phi(x)) + \epsilon\cdot\frac{\partial}{\partial\epsilon}R(\Phi_\epsilon(x))_{\epsilon=0} + O(\epsilon^2)$$

$$= x + \epsilon\cdot r(\Phi(x))\cdot\left[\Phi(x/10)-\Phi(x)\right] + O(\epsilon^2)$$

where $r(u)$ is the sparsity function $\frac{d}{du}R(u)$; see Hastings et al (1947) for the original definition. For our purposes we only note that $r(u)$ is easily obtained as

$$r(u) = \frac{d}{du}R(u) = \left[\frac{d\Phi(x)}{dx}\right]^{-1}_{x=R(u)} = \frac{1}{\phi(R(u))} .$$

In order to obtain an expression for $R_\epsilon(u)$ as $x+O(\epsilon^2)$, we introduce $H[R(\Phi_\epsilon(x))] = H(x)+O(\epsilon)$ with

$$H(x) = -r(\Phi(x))[\Phi(x/10)-\Phi(x)] \ .$$

This leads to

$$R(\Phi_\epsilon(x))+\epsilon H[R(\Phi_\epsilon(x))] = x+O(\epsilon^2)$$

or

$$x = R_\epsilon(u) = R(u) + \epsilon H[R(u)] + O(\epsilon^2) \ .$$

The first and second representing functions of X are now easily obtained as

$$R_0(u) = R(u)$$

$$R_1(u) = H[R(u)] = -r(u)\{\Phi(R(u)/10)-u\} \ .$$

$$= -\frac{\Phi(R(u)/10) - u}{\phi(R(u))}$$

The one-wild order statistic moments can now be numerically evaluated by substituting $R_0(u)$ and $R_1(u)$ into equation (3). Results of this are displayed in Table 1. We list the means and covariances of the one-wild order statistics for samples of size $n = 2(1)10$. For comparison purposes, the pure-Gaussian order-statistic moments are displayed in Table 2. As expected, the effects of contamination are most strongly evidenced in the extreme (or end) order statistics. More detailed tables have been computed by A. Bruce (1980).

REFERENCES

Bruce, Andrew G., 1980. "Tables of order statistics and their use in robust estimation," Senior Thesis, Princeton University (results will also appear as a technical report).

David, H. A., Kennedy, W. J., Knight, F. N. 1977. "Means, variances, and covariances of normal order-statistics in the presence of an outlier," Selected Tables of Mathematical Statistics, Volume 5, American Mathematical Society, Providence, R.I. pages ( ).

Hastings, C. Jr., Mosteller, F., Tukey, J. W. and Winsor, C.P. (1947). "Low moments for small samples: A comparitive study of order-statistics," Ann. Math. Statist. 18, 413-426.

Table 1

COVARIANCES FOR THE ONE-WILD GAUSSIAN SITUATION

N	I	E(X(I))	COV(X(I),X(J)) J= 1	2	3	4	5	6	7	8	9	10
2	1	-4.0093	34.4253	16.0746								
3	1	-4.2914	32.1211	2.2564	13.9034							
3	2	0.0000		0.9253								
4	1	-4.4379	30.9985	1.7175	1.5461	12.9093						
4	2	-0.4067		0.6407	0.4243							
5	1	-4.5359	30.2813	1.4869	1.1248	1.2626	12.2966					
5	2	-0.6413		0.5269	0.3312	0.2697						
5	3	0.0000			0.4295							
6	1	-4.6067	29.7637	1.3524	0.9503	0.8909	1.1052	11.8654				
6	2	-0.8043		0.4620	0.2849	0.2186	0.1978					
6	3	-0.2421			0.3471	0.2593						
7	1	-4.6632	29.3633	1.2619	0.8504	0.7402	0.7624	1.0032	11.5381			
7	2	-0.9281		0.4191	0.2556	0.1927	0.1619	0.1571				
7	3	-0.4136			0.3005	0.2229	0.1806					
7	4	0.0000				0.2782						
8	1	-4.7097	29.0391	1.1957	0.7840	0.6551	0.6253	0.6798	0.9306	11.2770		
8	2	-1.0272		0.3880	0.2349	0.1757	0.1444	0.1283	0.1312			
8	3	-0.5454			0.2698	0.1995	0.1603	0.1367				
8	4	-0.1737				0.2396	0.1913					
9	1	-4.7489	28.7682	1.1446	0.7359	0.5990	0.5485	0.5517	0.6218	0.8758	11.0616	
9	2	-1.1094		0.3642	0.2192	0.1632	0.1330	0.1151	0.1064	0.1134		
9	3	-0.6517			0.2476	0.1828	0.1465	0.1235	0.1091			
9	4	-0.3084				0.2141	0.1709	0.1426				
9	5	0.0000					0.2056					
10	1	-4.7829	28.5365	1.1036	0.6990	0.5587	0.4984	0.4805	0.5092	0.5784	0.8327	10.8792
10	2	-1.1795		0.3453	0.2069	0.1535	0.1245	0.1066	0.0956	0.0911	0.1004	
10	3	-0.7406			0.2306	0.1701	0.1362	0.1144	0.0997	0.0905		
10	4	-0.4179				0.1958	0.1563	0.1304	0.1122			
10	5	-0.1357					0.1932	0.1525				

Table 2

COVARIANCES FOR THE GAUSSIAN SITUATION

N	I	E(X(I))	COV(X(I),X(J)) J= 1	2	3	4	5	6	7	8	9	10
2	1	-0.5642	0.6817	0.3183								
3	1	-0.8463	0.5595	0.2757	0.1649							
3	2	0.0000		0.4487								
4	1	-1.0294	0.4917	0.2455	0.1580	0.1047						
4	2	-0.2970		0.3605	0.2359							
5	1	-1.1630	0.4475	0.2243	0.1481	0.1058	0.0742					
5	2	-0.4950		0.3115	0.2084	0.1499						
5	3	0.0000			0.2868							
6	1	-1.2672	0.4159	0.2085	0.1394	0.1024	0.0774	0.0563				
6	2	-0.6418		0.2796	0.1899	0.1397	0.1059					
6	3	-0.2015			0.2462	0.1833						
7	1	-1.3522	0.3919	0.1962	0.1321	0.0985	0.0766	0.0599	0.0448			
7	2	-0.7574		0.2567	0.1745	0.1307	0.1020	0.0800				
7	3	-0.3527			0.2197	0.1656	0.1296					
7	4	0.0000				0.2104						
8	1	-1.4236	0.3729	0.1863	0.1269	0.0947	0.0748	0.0602	0.0483	0.0368		
8	2	-0.8522		0.2394	0.1632	0.1233	0.0976	0.0787	0.0632			
8	3	-0.4728			0.2008	0.1524	0.1210	0.0978				
8	4	-0.1525				0.1872	0.1492					
9	1	-1.4850	0.3574	0.1781	0.1207	0.0913	0.0727	0.0595	0.0491	0.0401	0.0311	
9	2	-0.9323		0.2257	0.1541	0.1170	0.-034	0.0765	0.0632	0.0517		
9	3	-0.5720			0.1864	0.1421	0.1138	0.0934	0.0772			
9	4	-0.2745				0.1706	0.1370	0.1127				
9	5	0.0000					0.1661					
10	1	-1.5388	0.3443	0.1713	0.1163	0.0882	0.0707	0.0584	0.0489	0.0411	0.0340	0.0267
10	2	-1.0014		0.2145	0.1466	0.1117	0.0897	0.0742	0.0622	0.0523	0.0434	
10	3	-0.6561			0.1750	0.1338	0.1077	0.0892	0.0749	0.0630		
10	4	-0.3758				0.1579	0.1275	0.1058	0.0880			
10	5	-0.1227					0.1511	0.1256				

DATE
ILME